International Journal of Human-Computer Studies, in press

Evaluating a realistic agent in an advice-giving task.

Dianne. C Berry[1], Laurie T Butler[2], and Fiorella de Rosis[3]

1. Pro-Vice-Chancellors Office, University of Reading, Whiteknights House, Whiteknights, Reading, RG6 6AH. D.C.Berry@Reading.ac.uk.
   (Corresponding Author - Tel: 0118 378 7113. Fax: 0118 378 7424.
2. School of Psychology, University of Reading, Whitenights, Reading, RG6 6AL, UK. L.T.Butler@Reading.ac.uk
3. Department of Informatics, University of Bari, via Orbana, 70126, Bari, Italy. derosis@di.uniba.it.

Abstract

The aim of this study was to empirically evaluate an embodied conversational agent called GRETA in an effort to answer two main questions: 1) What are the benefits (and costs) of presenting information via an animated agent, with certain characteristics, in a 'persuasion' task, compared to other forms of display? 2) How important is it that emotional expressions are added in a way that is consistent with the content of the message, in animated agents? To address these questions, a positively framed healthy eating message was created which was variously presented via GRETA, a matched human actor, GRETA's voice only (no face), or as text only. Furthermore, versions of GRETA were created which displayed additional emotional facial expressions in a way that was either consistent or inconsistent with the content of the message. Overall, it was found that although GRETA received significantly higher ratings for helpfulness and likability, presenting the message via GRETA led to the poorest memory performance among users. Importantly, however, when GRETA's additional emotional expressions were consistent with the content of the verbal message, the negative effect on memory performance disappeared. Overall, the findings point to the importance of achieving consistency in animated agents.

Keywords: embodied animated agents; evaluation methods; believability; behaviour consistency

1. Introduction

Over the past ten years, there has been a considerable growth in interest in animated interface agents, that is computer generated characters that display life-like behaviours such as speech, emotions, gestures and eye, body and head movements (see Elliott & Brezezinski, 1998, for a review). Such agents differ from traditional interfaces in being able to act autonomously on behalf of the user, without the need for key or mouse-button clicks (Reeves and Nass, 1999). A long-term goal is to produce agents that can help out in a variety of contexts including education, e-commerce, tourist and service query systems, and entertainment (DeRosis, Pelachaud, Poggi et al, 2003; Johnson, Rickel and Lester, 2000). With this goal in mind, researchers have already developed a range of animated systems including PPP-Persona (Andre et al. 1998), REA (Cassell, 2001); see also Gratch and Masella (2001).

Broadly speaking, advocates of animated agents assume that agents serve to make computer systems more human-like, meaning that users can rely on well learned interaction skills (such as interpreting their partner's facial expressions or taking eye contact into account) which makes the interaction with the computer much smoother. There is also a general belief that agents have a motivational impact, making users more inclined to interact with a system (e.g., Lester & Stone, 1997). However, the use of animated agents has generated considerable debate among researchers (Norman, 1994; Wilson, 1997) with some arguing that this humanisation process may in fact hamper human-computer interaction. For example, it has been suggested that the presence of an eye-catching agent in a computer interface might actually result in a potential source of distraction (e.g. Walker, Sproull & Subramini, 1994). Additionally, the more realistic (i.e., human-like) the agent, the more likely it is that the user might apply 'social stereotypes', which can have undesirable consequences (e.g. Dehn & Van Mulken, 2000). As such, empirically establishing the impact (favourable or otherwise) of animated agents is of the utmost importance.

In informing this debate, a few studies have begun to evaluate the effects of animated agents in interfaces (e.g., Cassell and Thorisson, 1999; Lester et al. 1997; Sproull et al. 1996). For example, Takeuchi and Naito (1995) found that a virtual card matching game with an animated face was perceived as more entertaining than one where an arrow was used to visualise an opponent's moves. Similarly, Moundridou and Virvou (2002) reported that users enjoyed working with an algebra learning system more when fronted by an animated agent. However, empirical investigations of animated agents are still relatively rare and, as noted by Dehn and van Mulken (2000), have often produced conflicting results. Thus, whereas Sproull et al. (1996) reported that a virtual counsellor was rated higher on dimensions such as attractiveness and friendliness when presented as a textual output only, rather than with an additional face, Koda and Maes (1996) found that a poker player visualised by a face was more liked.

When one takes a closer look at the methods used in the different studies, however, it is not that surprising that results to date have been mixed, given that the researchers have tended to use a variety of different methods to evaluate very different

types of agent carrying out very different types of task. For example, many of the existing studies have used relatively simple 2-D characters, whereas a few have used more naturalistic 3-D characters displaying a much wider bandwidth of multimedia behaviours. Similarly, the type of task in which agents have been evaluated has varied widely (e.g., problem solving, game playing, educational/instructional), as has the specific role of the agent (e.g., friend, information provider, advisor). Finally, the measures used to assess performance have been even more diverse, with many studies using a range of different measures to assess users' subjective experience, others measuring user behaviour while interacting with the agent, and others still assessing their subsequent cognitive performance.

Given this diversity, and the inconsistent results to date, it does not seem particularly productive to ask whether the use of animated interface agents has beneficial effects or not. A more informative and fine-grained approach (see Dehn and Van Mulken, 2000, for similar ideas) might be to ask for which general classes of task (e.g., educational/instructional, game playing) are particular types of agent beneficial. By designing evaluation studies to address a more specific question, (such as whether a 2-D 'workman-like' character is helpful when used in a system for instructing people how to operate a mechanical device), researchers can start to build up a clearer overall pattern of findings in relation to establishing the costs and benefits of employing agents in interactive systems.

The aim of the present study is to contribute towards this broader goal by evaluating the effectiveness of an animated 3-D realistic character called GRETA (Pelachaud & Bilvi, 2003; Pelachaud et al, 2002). Given that the majority of existing studies have devised and assessed the effectiveness of agents in tasks or experimental settings that have involved providing instruction (such as learning to operate a device) or providing information (e.g. interacting with a travel agent, or playing a game), we chose persuasion as a class of task, where the purpose of providing information is to influence people's attitudes and encourage them to adapt their behaviour. Specifically, in our study, the domain selected involved persuading people to adopt healthier eating behaviours, as we thought this would be applicable to a broad section of the population and specifically to our intended participants (primarily University students). Furthermore, there is increasing awareness by health professionals and policy makers, within the UK and elsewhere, of the need to improve the communication of information about the importance of healthy eating.

We selected a realistic 3-D character, whose appearance was suited to the domain and participant population, in that she was youthful and healthy looking. The decision to use a female agent was based on the view, expressed by a number of our undergraduate students, that healthy eating is a predominantly 'female' topic. As Nass, Moon and Green (1997) reported that a female-voiced computer was rated by users as more informative about a stereotypically female topic, we felt that matching the gender of the agent with that of the topic was prudent.

Given that the purpose of the message involved persuasion as well as information provision, we employed a broad range of measures, including subjective ratings of the

quality of arguments and the likelihood that people would follow the advice, as well as measures of understanding and cognitive performance. We also evaluated people's perception of the agent itself.

A second issue addressed by the current study concerned consistency. When developing multimodal animated agents, a primary goal of designers is to achieve believability in the agent's behaviour. One key factor in achieving this is consistency (e.g., Loyall, 1997; Isbister & Nass, 2000). In human conversation such consistency occurs naturally, in that various forms of communication (e.g., hand gestures, facial expressions etc) are produced simultaneously with speech, and are normally consistent with the meaning being conveyed. Thus, it has been proposed that such consistency should be a feature of artificial agents. Ortony (2003), for example, has argued that for emotional agents to be believable there has to be more than just co-ordination of language and actions. Rather, the behaviours to be generated and the motivational states that subserve them have to have some consistency. However, modelling such multi-modal behaviour in the case of animated agents represents a considerable technical challenge (see e.g., Cassell et al. 2000). In informing on this process, one question that is currently unclear is the extent to which failure to achieve consistency *actually* results in costs to the effectiveness of an agent. Thus, an additional aim of the present study is to use GRETA to investigate empirically the importance of achieving consistency in multimodal agents. Specifically, we examined the effects of adding emotional expression in either a consistent or an inconsistent way to assess whether or not this was beneficial. In a previous study, Isbister and Nass (2000) illustrated the importance of consistency, finding that users preferred computer characters that displayed consistent verbal and non-verbal cues (e.g., extrovert language and posture). Therefore, while the specific behaviors being manipulated in the present study are rather different, it should be expected that GRETA will be most effective where emotional expression is added in a consistent way.

The objectives of the current study are therefore two-fold: first, to establish the benefits (and costs) of presenting information via an animated agent (possessing certain characteristics) in a 'persuasion' task and, second, to assess the effects of adding emotional expression to the agent in either a consistent or inconsistent way. For comparison purposes, we selected three additional presentation modes; a (matched) human actor condition, a voice (GRETA's) condition and a written text condition. The use of matched control conditions and a broad range of measures in this study, is in line with Dehn and Van Mulken's (2000) comment that the choice of control group and dependent variable are crucial when evaluating agent effectiveness. Interestingly, Moreno, Mayer, Spires and Lester (2001) recently examined the role of agent presentation mode in encouraging learning in a computer-based environment. Briefly, participants retained more information, showed greater knowledge transfer to new problems, and were more interested in the computer based lessons overall when the materials were communicated via speech rather than as on screen text. Furthermore, the visual presence of an agent itself, whether animated or human, neither enhanced nor impaired learning. On this basis we might predict an overall disadvantage, in terms of presenting our healthy eating dialogue, for our text only condition. However, it is

important to note that there are crucial differences between the present study and the one reported by Moreno et al (2001), not least in terms of the choice of agent employed (realistic 3-D character capable of complex facial expressions versus cartoon figure) and the task domain (persuasion versus educational task). Therefore, we must be cautious when drawing parallels between the two studies.

2. Method

2.1 Participants

One hundred and forty four participants from the University of Reading participated in the experiment, for which they received either course credit or a small cash payment of £3. Twenty-four participants were randomly assigned to one of six presentation conditions.

2.2 Design

The main experiment consisted of a single independent variable comprising four conditions (presentation type: GRETA [neutral] vs. Human [neutral] vs. GRETA voice only vs. text only). Presentation type was manipulated between participants. Additionally, to examine the role of consistency of behaviour in an animated agent, two further conditions were included GRETA [consistent] vs. GRETA [inconsistent].

2.3 Materials

A message was created which was intended to encourage healthy eating behaviour (see Appendix 1). This consisted of a series of short statements that emphasized the positive consequences of engaging in healthy eating (e.g., 'As far as vitamins are concerned, research has shown that eating the recommended levels of vitamin A and C can have beneficial effects for your appearance and health'). The statements were framed in a positive way to permit versions of the message to be created whereby the animated agent's added emotional expressions were either consistent (smiling) or inconsistent (concerned) with the content of the message itself (see below).

In producing the materials for the experimental conditions, the healthy eating message was manipulated in various ways. First, a text only condition was created, which involved printing and displaying the healthy eating message in size 12 Arial font on a sheet of A4 paper (as displayed in Appendix 1). Next, for the animated agent conditions, three video clips of GRETA (head only) delivering the healthy eating message were produced jointly by researchers working at the Universities of Roma La Sapienza, Edinburgh and Bari (see figure 1). The voice of GRETA was produced using a text to speech synthesizer, called FESTIVAL[i]. Given that the aim of the current study was to evaluate the role of consistency in facial expressions, no emotional cues (e.g., higher pitch, faster speech to indicate a cheerful tone) were added to GRETA's voice itself. In the consistent version, GRETA smiled at contextually appropriate points whilst delivering the message (e.g., when saying 'can lead to good skin and hair') whereas in the inconsistent version, GRETA displayed a concerned facial expression (e.g., eyes looking downwards, arching of the eyebrows) at the same points. In a final neutral version,

GRETA relayed the message without displaying either type of expression. It should be noted that, in all three conditions, GRETA displayed a variety of natural behaviours such as eye gazing and head turning which were unrelated to consistency.

Figures 1, 2 and 3 about here

Expressions were produced by giving, to the agent player, input in the form of a XML file whose tag were defined according to the APML markup language (De Carolis et al, 2002). Appendix 2 contains an example of part of this file, which shows where 'affect' tags (in this case, of 'small joy') are incorporated, in the 'consistent' version of the message. In the 'inconsistent' version, a tag of 'small sadness is associated with the same text portions, while in the 'neutral' version no affect tag is introduced.

The video clips were shown via a Windows Media Player on a 1Ghz PC with a 17" monitor. The size of the display was 20 cm x 15 cm and was viewed at a distance of 50 cm by participants. Each clip lasted approximately two and a half minutes. Finally, in producing the voice only condition, the neutral GRETA video clip was simply played to participants whilst the Windows media player was minimized on screen. As such, the voice only condition was identical to the neutral condition (i.e. produced via FESTIVAL) except that GRETA was not visually present on the screen.

In addition to the five presentation conditions described above, a sixth condition was produced at the University of Reading using a human actor. In choosing a suitable actor, the aim was to find someone who was as closely matched as possible to GRETA on key dimensions such as age, sex, and visual appearance. To this end, a female actor was selected who broadly matched each of these criteria (i.e., a young fit-looking woman with short dark hair). A video clip of the actor presenting the healthy eating message, in conjunction with neutral expressions, was filmed using a digital video camera. To ensure that the actor's expressions matched those of GRETA as closely as possible, the actor first studied the relevant video clips of GRETA. Additionally, when filming, the actor read from an auto-cue that contained mark ups at the precise points at which GRETA displayed facial expressions. Once recorded, the video clip was transferred to videotape. Given that we did not want participants to mistake the human actor for a computer agent, the clip was displayed via a video recorder on a standard 21" television screen rather than on a computer monitor. Had the clip been displayed via the computer monitor, there was a concern that participants would have treated the actor as a photo-realistic animated agent. The video clip itself was viewed at a range of approximately 1.5 metres, to compensate for the larger absolute image size compared to GRETA and lasted for approximately 3 minutes.

For each of the presentation conditions, a short introductory message was produced. The message for the text only condition read as follows: 'The Food Standards Agency is an independent food safety watchdog that was set up in 2000 to protect the public's health and consumer interests in relation to food. One of its major aims is to

encourage and help people to eat more healthily. The aim of this study is to evaluate one of the messages provided by the Agency. We would like you first to answer some questions about yourself to ascertain your general views on healthy eating. We would then like you to read a piece of text containing a healthy eating message produced by the Agency. Finally, we would like you to answer some questions about the message'. For the other conditions, minor wording changes were made to the penultimate sentence as appropriate. However in all cases, the introductory explanation was presented twice in succession in the same form as the healthy eating message itself (i.e., via GRETA, human, voice only or as text only). This was done to familiarize participants with the particular form of presentation to which they were assigned, before the crucial healthy eating message was delivered. It should be noted that for conditions involving GRETA, the introductory message was always presented by the neutral version of the agent.

For the purposes of evaluating user and agent performance, two paper-based questionnaires were constructed. First, a *pre-test questionnaire* (Appendix 2) was designed which contained items on level of healthy eating knowledge, and existing dietary behaviour. Participants responded by means of unipolar six point scales. The questionnaire was intended to ensure that any post-test differences obtained between participants assigned to the various healthy eating message presentation conditions were not due to pre-experimental differences in healthy eating knowledge. Second, a *post-test questionnaire* (Appendix 3) was constructed to measure the impact of the healthy eating message on participants' subsequent attitudes and behaviour. As noted earlier, this comprised a range of different measures. To assess the effectiveness of the 'information provision' aspect of the message, we included ratings of satisfaction and ease of understanding (both on six-point unipolar Likert scales). In addition, a sequence of explicit memory probes was devised to test participants' memory for the message. This consisted of both multiple-choice (e.g., According to the healthy eating message a recommended portion of fruit and vegetables is 40g, 25g, 90g, 70g, 80g, or 65g) and true/false questions (Vitamin A is important for good hearing - True/False). To assess the 'persuasive' aspect of the message, we included ratings of perceived persuasiveness, convincingness, and trustworthiness of the message, as well as of the quality of the evidence and arguments (all six-point unipolar Likert scales). We also assessed the likelihood that users would follow the advice given in the message (on a six point unipolar Likert scale). Finally, a set of items was created to measure participants' perceptions towards the agent (GRETA and human) themselves. This required participants to rate the agent, along the following dimensions; helpfulness, intelligence, credibility, likability, reliability, competency and believability. All ratings were made using six-point unipolar Likert scales.

2.4 Procedure

All participants were tested individually and, upon entering the testing room, first received a short explanation describing the purpose of the experiment. This was presented twice, via GRETA, the human agent, voice only, or as text (to familiarize participants with their assigned form of presentation). Participants in the GRETA and

voice only conditions received the introductory message via a computer monitor, whereas those in the human agent conditions viewed a television screen. Finally, participants in the text only condition read the explanation (silently) from a sheet of paper in front of them.

All participants were then presented with the written pre-test questionnaire. This took no longer than five minutes to complete. Subsequently, all participants received the healthy eating message. In the text only condition, participants were instructed to read the message (silently) once only at their own pace, after which the message was removed. For the remaining presentation conditions, the message was presented once in the same medium as the introductory explanation. Finally, all participants were given the post-test written questionnaire and were asked to complete it as accurately and honestly as they could. The questionnaire took no more than ten minutes to complete, after which participants were fully debriefed.

3. Results

3.1 Comparing different forms of presentation

Initially, the data for the neutral version of GRETA was compared with that from the neutral human agent, voice only, and text only conditions. This permitted a clean assessment (i.e., without the potential confound of additional facial expressions) of just how well GRETA performed in comparison to alternative methods of message presentation.

An analysis of the pre-test questionnaire data using a one way ANOVA revealed no statistically significant differences between the four conditions in terms of levels of healthy eating knowledge, $F(3, 92) = .352$, $p > 0.1$, or existing dietary behaviour, $F(3, 92) = 1.461$, $p > 0.1$. This was important as it meant that we could be confident that any post intervention differences between the groups were not due to pre experimental differences in healthy eating knowledge between the participants.

The data from the post-test questionnaires were analysed using one way ANOVAs. Initially, we examined whether the effectiveness of the healthy eating message varied across the four presentation conditions. As can be seen in figure 2, participants' mean ratings of their likelihood of following the recommendations for healthy eating were very similar across the GRETA, human, voice and text only conditions. An ANOVA confirmed that there were no significant differences between the various conditions, $F(3, 92) = .788$, $p > 0.1$. However, there were some differences across the four conditions in terms of participants' perceptions towards the healthy eating message. First, there was a significant difference in terms of ease of understanding the message, $F(3, 92) = 6.365$, $p = .001$. Post hoc tests showed that the text version of the message was rated as being easier to understand than either the voice $t(46) = 3.597$, $p = .001$, GRETA, $t(46) = 3.871$, $p = .0001$, or human, $t(46) = 1.984$, $p = .05$, conditions.

There was also a significant effect of trustworthiness of the message, $F(3, 92) = 3.394$, $p = .02$, such that ratings were higher for voice, $t(46) = 2.598$, $p = .01$, text, $t(46) =$

2.530, p=.02, and GRETA, t(46) = 3.054, p=.004, conditions compared to the human condition. No other differences approached significance.

Figure 4 about here

In terms of perceptions of the various agents themselves, comparisons were made between the Greta, Voice, and Human conditions, (see Figure 3). The results showed a significant effect of helpfulness, F(2, 69) = 4.459, p=.05, and an effect of likability, F(2, 69) = 3.785, p=.05. Post hoc tests showed that ratings of helpfulness were higher for both human, t(46)= 2.877, p=.006, and GRETA conditions, t(46)=2.266, p=.03, compared to the voice only condition. A similar pattern was observed for likability in that the voice only condition was significantly less liked than the human agent condition, t(46) = 2.846, p=.007. However, the difference between GRETA and the voice only condition did not quite achieve significance, t(46)=1.923, p=.06. Overall though, the data demonstrates that the presence of a face (human or animated) does confer some clear advantages in terms of ratings of the agent.

Figure 5 about here

Finally, participants' overall memory performance was examined. The proportion of correct answers given by participants assigned to GRETA neutral, human neutral, text only and voice only were .71, .79, .78, and .80 respectively (see figure 4). This indicates that retention of the healthy eating message was poorest for participants assigned to the neutral GRETA condition. This was confirmed by a one way anova which revealed a significant difference between the four groups, F(3, 90) = 3.678, p=0.015.[ii] Post hoc tests showed that the difference between the GRETA neutral condition was significant in the case of the human, t(45) = 2.91, p = .006, text, t(46) = 2.266, p=.03, and voice only control, t(45) = 2.951, p=.005.

Figure 6 about here

3.2 The importance of consistency of behaviour in an animated agent

To evaluate the effect of adding emotional expression to the agent, in either a consistent or inconsistent way, a further analysis was conducted comparing the neutral, consistent and inconsistent versions of GRETA directly.

Firstly, the pre-test questionnaire data was analysed. Again, an ANOVA showed no significant difference between the three presentation conditions in terms of existing

levels of healthy eating knowledge, $F(2, 69) = .582$, $p>0.1$, or existing dietary behaviour, $F(2, 69) = .588$, $p>0.1$.

Turning to the post-test questionnaire (see Figure 5), there were no significant differences between the three groups in terms of behavioural intentions to engage in healthy eating, $F(2, 69) = .987$, $p >0.1$. However, there was a marginally significant effect of quality of evidence, $F(2, 69) = 2.772$, $p=.07$. This showed that participants viewed the healthy eating evidence most favourably when presented by the neutral version of GRETA, significantly so when compared to the consistent version, $t(46) = 2.356$, $p=.02$ A similar pattern was observed for convincingness, $F(2, 69) = 3.885$, $p=.03$, such that the evidence was perceived most favourably when presented by the neutral version of the animated agent rather than by consistent, $t(46) =2.299$, $p=.02$, or inconsistent versions, $t(46) = 2.663$, $p=.01$. This was also the case for trustworthiness of the evidence, $F(2, 69) = 3.85$, $p=.03$, where again the neutral version of GRETA outperformed versions that additionally incorporated consistent, $t(46) = 1.951$, $p=.05$, and inconsistent, $t(46) = 2.930$, $p=.005$, facial expressions.

Figure 7 about here

There were no significant differences in participants' perceptions towards the three versions of GRETA on any of the dimensions. However, cognitive performance did differ. Specifically, memory for the healthy message was greatest when presented by the consistent version of GRETA (.76) compared to the neutral (.71) and inconsistent (.71) versions (see figure 4). Post hoc tests revealed that the difference in memory performance was significant for the consistent and neutral groups, $t(46) = 2.058$, $p=.05$, and for the consistent and inconsistent groups, $t(46) = 1.984$, $p=.05$. Thus, adding emotional expressions in a consistent way into animated agent design does seem to provide a clear operational advantage.

4. Discussion

The present study revealed four main findings. First, participants performed more poorly in terms of cognitive performance when the healthy eating message was presented by GRETA (neutral) than via the other media. Second, and in contrast, there were some positive consequences of employing an animated agent in terms of participants' ratings of helpfulness and liking (and also trustworthiness of the message). Third, the inclusion of additional facial expressions (consistent or inconsistent) resulted in lower ratings of the message itself (including trustworthiness) when compared to the neutral version of GRETA . Fourth, and most importantly, when the emotional expressions were added in a consistent way the resulting cognitive performance was comparable to any other method of presentation.

In terms of interpreting the current findings, an initial question is why the neutral version of GRETA resulted in poorer user performance compared to the human, text and

voice only conditions. It was not the case that GRETA was more difficult to understand as there was no significant difference in ease of understanding between the GRETA condition and the human and voice conditions. Similarly, the effects cannot be attributed to differences in presentation speed. Compared to GRETA, the message was delivered over a longer period of time in both the human and text only conditions, a fact reflected in users' subsequent ease of understanding ratings. However, both message speed and comprehensibility were the same in the case of the GRETA neutral and voice only conditions. Despite this, the cognitive performance of users assigned to the latter group was substantially better. Given that the only difference between the two groups was the presence of GRETA's face, a more likely explanation is that the presence of the face distracted participants. Interestingly, and in support of the current findings, a study by Wright, Milroy and Lickorish (1999) has shown that animated graphics may have a detrimental effect on user performance in terms of hampering text retention. Given limited cognitive resources, it is well established in cognitive psychology that the presence of a distraction during encoding can lead to poorer subsequent memory for target information (e.g., Banbury & Berry, 1997, 1998). One unresolved question though, is why the presence of a human face did not result in similarly poor cognitive performance. One possible explanation is that users perceive the presence of an animated head as more novel and deserving of attention than a human head. We return to this issue in a later section.

The fact that the GRETA neutral condition resulted in the poorest cognitive performance, compared to the other presentation forms, could imply that this kind of character is inappropriate for 'information giving' tasks, where the information is sensitive and the user needs to remember it correctly. However, such a character might still be appropriate for less sensitive tasks, such as programme presentation on digital TV, in which the information is provided within the context of entertainment. It should also be remembered that there were some positive consequences of employing GRETA compared to the human condition (in terms of trustworthiness) and the voice only condition (in terms of helpfulness and likeability). The former may reflect the tendency to view machines as less likely to lie than human beings, whereas the latter may demonstrate the human tendency to see computers as teammates (Nass, 1996). Thus, this type of character might be more acceptable for tasks (like e-learning) in which persuasion and the establishment of a trust relationship with the tutor may benefit learning (particularly for young students).

One further comment with respect to cognitive performance is that our text only condition produced user performance levels that were at least as good as those from any of our other conditions (all of which involved some form of auditory presentation). This is in direct contrast to Moreno et al. (2001) who showed an advantage for auditory over text based presentation of information, both in terms of memory retention, knowledge transfer, and user ratings of the lesson itself. One possible explanation for the discrepant findings is that text was presented via paper in the current study, rather than on via a computer screen as in the Moreno et al. study. While the literature is somewhat inconsistent, there is some evidence suggestive of advantages for paper over screen

processing of text (see Dillon, 1992, for review). Additionally, the text condition in the current study permitted a greater degree of user control than the other conditions in the sense that it was self-paced. This is reflected in the ease of understanding ratings given by users, and could have had the effect of 'canceling out' of any auditory small advantage that was present.

One puzzling aspect is why the message was rated as less trustworthy when presented by the consistent and inconsistent versions of GRETA than by the neutral version. This may be a result of the artificiality of the added emotion. Despite the technical achievements that were necessary in creating GRETA, embodying animated agents with realistic emotions is still at an early stage. The most important finding in relation to the issue of adding emotional expression in a consistent or inconsistent way, though, is that cognitive performance is actually as good as any form of presentation when presented by a 'consistent version' of GRETA. One possible explanation for this is that the presence of contextually relevant facial cues aids encoding of the message itself. People may become more engaged with the interaction and this, in turn, may offset the distracting influence of the animated face itself. Clearly, the mechanisms involved here will require investigation in future studies. Finally, it is worth noting that, although achieving consistency in behaviour is clearly important in terms of user performance, no version of GRETA was actually better than the alternative forms of presentation. The current study has merely shown that when GRETA was at her best she was no worse than any other medium.

The above findings and preliminary conclusions need to be interpreted in the light of the constraints and limitations of the current study. First, we used healthy eating as our message topic, whereas it may be that GRETA would be better suited to other domains. Perhaps her youthful physical appearance would make her more suited to providing advice in domains that are more difficult to initially engage users, such as physics or chemistry teaching aids, or to taking the role of friend in a game playing situation. It may also be that GRETA's youthful appearance was not appropriate for the role persuading users to change their health-related behaviours. Users may have preferred a more 'professional looking' character, although it should be remembered that participants rated themselves as just as likely to follow the recommendations for health eating when presented by GRETA as when presented by the other media. In general, however, it would have been preferable to use more than one agent type in the study to ensure that our findings are not the result of GRETA's specific appearance. As Dehn and Van Mulken (2000) note, this is a limitation that is common in virtually all empirical studies in this area.

Second, GRETA may have had a larger positive impact on performance if users had been given opportunity to become more familiar with her over time. We were not able to examine this in the present study as we only used a single, relatively short, message. Another limitation is that we measured retention of the message (using cognitive memory measures) directly after presentation of the message. It would be interesting to see what happens after a delay, by which time message retention will be much poorer. For

example, will participants with better memory for the message evaluate its content more positively?

Finally, GRETA did not have an opportunity to 'communicate' in an interactive way with users in the current project. The interaction was limited to a video clip, which the user watched passively. Although a necessary first step before moving on to more elaborate forms of interaction, it is likely that more advantages of employing animated agents will emerge as the interaction becomes more demanding and complex.

In conclusion, despite these limitations, the present study has made some contribution towards identifying the tasks for which a realistic agent like GRETA might be appropriately employed. In addition, our findings suggest that careful consideration needs to be given to determining how emotional expression is added to animated agents. While cartoons may manifest their emotions through paradoxical behaviours, 3-D realistic agents that pretend to emulate humans should show natural expressions if they are to appear convincing, and engender trust in the user.

Acknowledgments

We would like to thank Paula Brown, Krissy Wilson and Sofia Boucas for their assistance in collecting the data reported in this study. The work was conducted as part of the MAGICSTER project supported by EC grant IST-1999-29078. We are grateful to our project collaborators, particularly Catherine Pelachaud, for their contributions to the work.

Appendix 1

Positively framed healthy eating message

The Food Standards Agency in the UK, and Departments of Health throughout the Western World, recommend that people should aim to eat at least five portions of fruit and vegetables a day. These can be fresh, frozen, tinned, dried or juiced, although eating them raw or lightly cooked is best. Potatoes should not be counted because they contain starch. A recommended portion is roughly 80g, or for example, an apple, 2-3 tablespoons of frozen vegetables, or 1 glass of fruit juice. Children can be given smaller portions, according to their appetites. Although the recommended amounts may seem a lot, fruit and vegetables are good sources of vitamins (particularly A and C), minerals, and fibre.

As far as vitamins are concerned, research has shown that eating the recommended levels of vitamin A and C can have beneficial effects for you appearance and health. These vitamins not only cleanse the blood but are important for growth and the repair of body tissues.

In terms of your appearance, eating foods which are rich in vitamin A and C can lead to good skin and hair, as vitamin C stimulates, tonifies and rehydrates them. Eating these vitamins will also result in healthy gums and strong teeth.

Diets that are rich in vitamin A and C are also good for your health. They give resistance against colds, flu and other infections, and generally result in higher energy levels. They are also associated with a healthy heart and stomach.

Consuming extra amounts of these vitamins has been shown to relieve asthma, improve blood sugar control, and promote healing. Vitamin A is essential for good vision, particularly in dim lighting conditions.

Many fruit and vegetables also contain minerals such as calcium, which are important for strong bones and teeth, effective muscles, and a well-functioning nervous system.

They are also a good source of fibre. Eating sufficient quantities of fibre is important for a healthy heart and effective digestive system, and can improve dental well-being.

Appendix 2

==========


```
<?xml version="1.0"?>
<apml>
.....
<performative type="inform">
<theme affect="small_joy">
In terms of your
<emphasis x-pitchaccent="LplusHstar">appearance, </emphasis><boundary type="LH"/>
</theme>
</performative>
<performative type="small_incite">
<theme affect="small_joy">
<emphasis x-pitchaccent="LplusHstar">eating </emphasis>
foods which are rich in vitamin
<emphasis x-pitchaccent="LplusHstar">A </emphasis>
and
<emphasis x-pitchaccent="LplusHstar">C</emphasis> <boundary type="LH"/>
</theme>
<rheme affect="small_joy">
can lead to
<emphasis adjectival="big" x-pitchaccent="Hstar">good </emphasis>
<emphasis x-pitchaccent="Hstar">skin </emphasis>
and
<emphasis x-pitchaccent="Hstar">hair, </emphasis><boundary type="LH"/>
</rheme>
<theme affect="small_joy">
<emphasis x-pitchaccent="LplusHstar">as </emphasis>
vitamin
<emphasis x-pitchaccent="LplusHstar">C </emphasis><boundary type="LH"/>
</theme>
<rheme affect="small_joy">
<emphasis x-pitchaccent="Hstar">stimulates, </emphasis>
<emphasis x-pitchaccent="Hstar">tonifies </emphasis>
and
<emphasis x-pitchaccent="Hstar">rehydrates </emphasis>
them.<boundary type="LL"/>
</rheme>
</performative>
…….
</apml>
```


==========

Appendix 3 : Pre-Test Questionnaire

Questionnaire 1

Some questions about you…

1) How much would you say you know about healthy eating?

1 2 3 4 5 6

Very little knowledge Considerable knowledge

2) To what extent do you consider yourself to be a healthy eater?

1 2 3 4 5 6

Not at all Extremely

3) To what extent do you try to eat a diet that is rich in vitamins and minerals?

1 2 3 4 5 6

Never Always

4) What do you see as the main benefits (if any) **and** costs (if any) of eating a diet that is rich in vitamins and minerals?

Benefits:

Costs:

Age __ **Sex** M / F (please circle one)

End of questionnaire

Appendix 4 : Post-Test Questionnaire

Questionnaire 2

1) On the basis of what you have just heard how likely are you to follow the recommendations for healthy eating in the future?

1	2	3	4	5	6
Not at all likely					Extremely likely

2) How satisfied were you with the information contained in the message?

1	2	3	4	5	6
Not at all satisfied					Extremely satisfied

3) How easy did you think that the message was to understand?

1	2	3	4	5	6
Not at all easy					Extremely easy

4) How persuasive did you think the healthy eating message was?

1	2	3	4	5	6
Not at all persuasive					Extremely persuasive

5) How convincing did you think the healthy eating message was?

1	2	3	4	5	6
Not at all					Extremely

6) Overall, how would you rate the quality of the arguments contained in this message?

1	2	3	4	5	6

Not at all good | Extremely good

7) How trustworthy did you think the content of the message was?

1 2 3 4 5 6

Not at all trustworthy | Extremely trustworthy

8) Please circle any of the following words that appeared in the message

a) sodium calcium magnesium iron zinc

b) teeth hearing colds diarrhea headaches

c) liver spleen heart bowels stomach

d) Vitamin A Vitamin D Vitamin E Vitamin C vitamin B

9) How many portions of fruit and vegetables a day does the Food Standards Agency in the UK recommend that people should eat?

four seven two five three six (circle one)

10) According to the message a recommended portion of fruit or vegetables is……..

40g 25g 90g 70g 80g 65g (circle one)

11) On the basis of what you heard in the message, please mark the following statements as being True or False

a) Vitamin A is necessary for growth and repair of body tissues

True False

b) Vitamin A is important for good hearing

True False

c) Vitamin A and C are necessary for an effective digestive system

True False

d) A shortage of minerals such as calcium can have detrimental effects for our muscles and nervous system

True False

e) Fibre helps to cleanse the blood

True False

f) Vitamin A and C are good for our gums and teeth

True False

12) For each word below, please indicate how well it describes the **computer character** that you have just seen. Note that you are now evaluating the character, **NOT** the message.

a) Helpful

1	2	3	4	5	6
Not at all					Extremely

b) Intelligent

1	2	3	4	5	6
Not at all					Extremely

c) **Credible**

1	2	3	4	5	6
Not at all					Extremely

d) Likable

1	2	3	4	5	6
Not at all					Extremely

e) Reliable

1	2	3	4	5	6
Not at all					Extremely

f) Believable

1	2	3	4	5	6
Not at all					Extremely

g) Competent

1	2	3	4	5	6
Not at all					Extremely

End of questionnaire.

References

Andre, E., Rist, T., Muller, J., 1998. Integrating reactive and scripted behaviors in a life like presentation agent. Autonomous. Agents 98. 261-8.

Banbury, S., Berry, D.C., 1997. Habituation and dishabituation to speech and office noise. Journal of Experimental Psychology: Applied, 3, 181-195.

Banbury, S., Berry, D.C., 1998. Disruption of office-related tasks by speech and office noise. British Journal of Psychology, 89, 499-517.

Cassell, J., 2001. Embodied Conversational Agents. AI Magazine, Winter 2001, 67-83.

Cassell, J., Bickmore, T., Campbell, L., Vilhjálmsson, H., Yan, H., 2000. Human Conversation as a System Framework: Designing Embodied Conversational Agents, in Cassell, J. et al. (eds.), Embodied Conversational Agents, pp. 29-63. Cambridge, MA: MIT Press

Cassell, J., Bickmore, T., Campbell, L., Vilhjálmsson, H., Yan, H., 2001. More than just a pretty face: conversational protocols and the affordances of embodiment. Knowledge-Based Systems, 14, 55-64.

Cassell, J., Thorisson, K. R., 1999. The power of a nod and a glance: Envelope versus emotional feedback in animated conversational agents. Applied AI, 13, 519-538.

De Carolis, B., Pelachaud, C., Poggi, I. & Steedman, M., 2002. APML, a markup language for believable behavior generation. In Prendinger, H. and Ishizuka, M. (Eds). Life-like characters. Tools, affective functions and applications. Springer.

Dehn, D.M., van Mulken, S., 2000. The impact of animated interface research: A Review of empirical research. International Journal of Human-Computer studies, 52, 1-22.

De Rosis, F., Pelachaud, C., Poggi, I., Carofiglio, V., De Carolis, B., 2003. From Greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent. International Journal of Human-Computer Studies. Special Issue on "Applications of Affective Computing in HCI". 59, 81-118.

Dillon, A., 1992. Reading from paper versus screens: A critical review of the empirical literature. Ergonomics, 35, 1297-1326.

Elliott, C., Brezezinski, J.,1998. Autonomous Agents as Synthetic Characters. AI Magazine, Summer, 13-30.

Gratch, J., Marsella, S., 2001. Tears and Fears: Modeling emotions and emotional behaviors in synthetic agents. Proceedings of the 5th International Conference on Autonomous Agents, Montreal, Canada,.

Isbister, K., Nass, C., 2000. Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. International Journal of human-computer studies, 53 (2): 251-267.

Johnson, W., Rickel, J., Lester, J., 2000. Animated pedagogical agents: Face-to-face interaction in interactive learning environments. International

Journal of Artificial Intelligence in Education, 11, 47-78.

Koda, T., Mayes, P., 1996. Agents with faces: the effects of personification of agents. Fifth IEEE International Workshop on Robot and Human Communication.

Lester, J. C., Stone, B. A., 1997. Increasing believability in animated pedagogical agents. Proceedings of 1st International Conference on Autonomous Agents, 16-21.

Lester, J., Converse, S., Stone, B., Kahler, S., Barlow, T., 1997. Animated pedagogical agents and problem-solving effectiveness: A large-scale empirical evaluation. In Boulay & Mizoguchi (Eds.) Proceedings of the 8th World Conference on AI in intelligence. IOS press.

Loyall, A.B., 1997. Believable agents: Building Interactive Personalities. PhD thesis, Carnegie Mellon University.

Moreno, R., Mayer, R. E., Spires, H., Lester, J., 2001. The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? Cognition and Instruction, 19, 177-213.

Moundridou, M., Virvou, M., 2002. Evaluating the persona effect of an interface agent in a tutoring system. Journal of Computer Assisted Learning, 18, 253-261.

Nass, C., Fogg, B J., Moon, Y., 1996. Can computers be teammates? International Journal of Human-Computer Studies, 45, 669-678.

Nass, C., Moon, Y., Green, N., 1997. Are computers gender-neutral? Gender stereotypic responses to computers. Journal of Applied Social Psychology, 27(10), 864-876.

Nass, C., Moon, Y., Fogg, B J., Reeves, B., Dryer C D., 1995. Can computer personalities be human personalities? International Journal of Human-Computer Studies, 43, 223-239.

Norman, D.A., 1994. How might people interact with agents. Communications of the ACM, 37, 68-71.

Ortony, A., 2003. On making believable emotional agents believable. In R. Trappl, P. Petta and S.Payr (Eds): Emotions in humans and artifacts. MIT Press, Cambridge, MA.

Pelachaud, C. Bilvi, M., 2003. Computational model of believable conversational agents. In MP Huget (Ed): Communication in Multiagent systems: Background, current trends and future. Springer Verlag: Lecture Notes in Computer Science Vol. 2650. Pp 300-17.

Pelachaud, C., Carofiglio, V, De Carolis, B., De Rosis, F. Poggi, I., 2002. Embodied contextual agent in information delivery application. First International Joint Conference on Automatic Agents and Multi-Agent systems. Bologna. July 2002.

Reeves, B., Nass, C., 1999. The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. Cambridge University Press, New York, NY.

Sproull, L., Subramani, M., Kiesler, S., Walker, J.H., Waters, K., 1996. When

the interface is a face. Human-Computer Interaction, 11, 97-124.
Takeuchi, A., Naito, T., 1995. Situated Facial Displays: Towards Social Interaction, Proceedings of the 1995 ACM Conference on Human Factors in Computing Systems (CHI `95), New York: ACM Press, 450-455.
Walker, J.H., Sproull, J. Subramani, R., 1994. Using a Human Face in the Interface, Proceedings of the 1994 ACM Conference on Human Factors in Computing Systems (CHI `94), New York: ACM Press, 85-91.
Wilson, M., 1997., Metaphor to personality: the role of animation in intelligent interface agents. Proc. IJCAI-97 Workshop on Animated Interface Agents: Making them intelligent.
Wright, P., Milroy, R., Lickorish, A., 1999. Static and animated graphics in learning from interactive texts. European Journal of Psychology Education, 14, 203-224.

Figure Legends

Figure 1: Still images of GRETA with a smiling expression

Figure 2: Still images of GRETA with a concerned expression

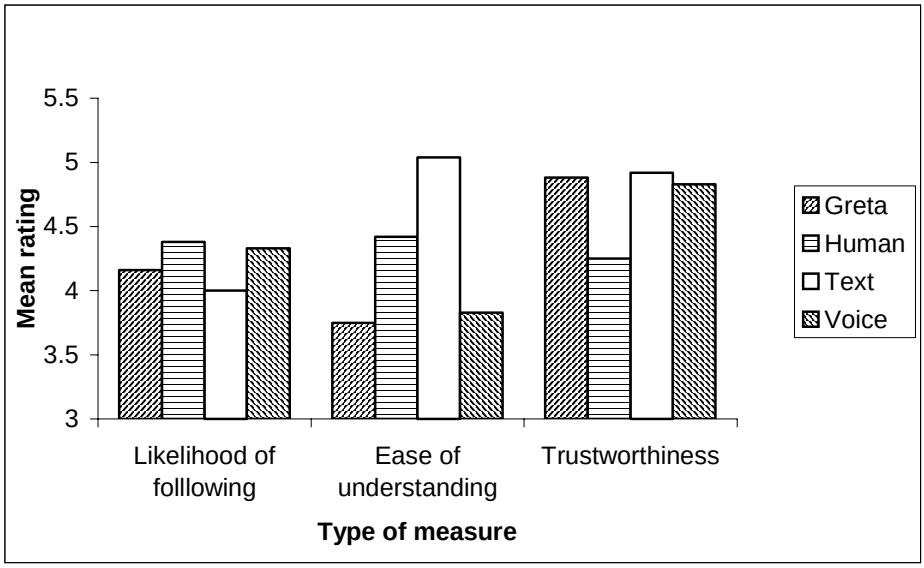
Figure 3: Still images of GRETA with a neutral expression

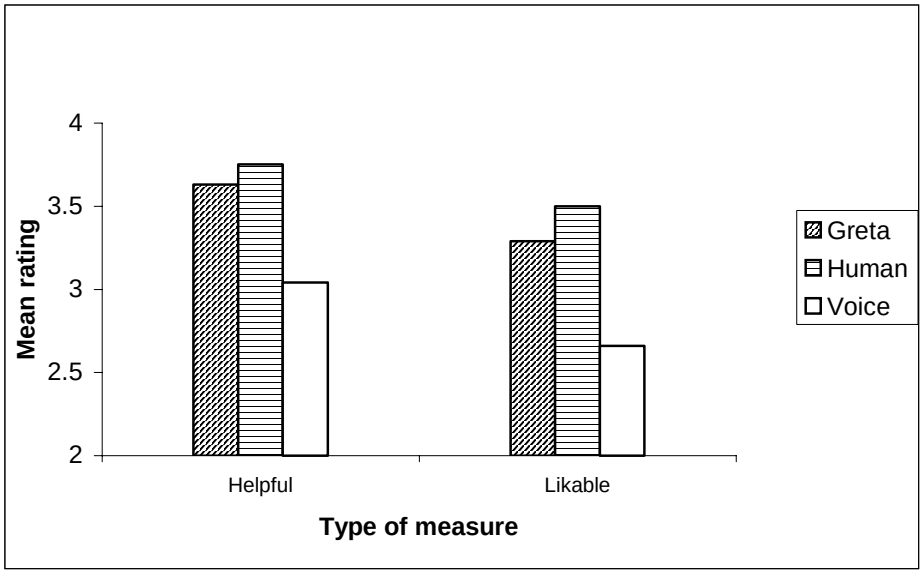
Figure 4: Perceptions of participants towards the healthy eating message: comparison of Greta, human, text and voice forms of presentation

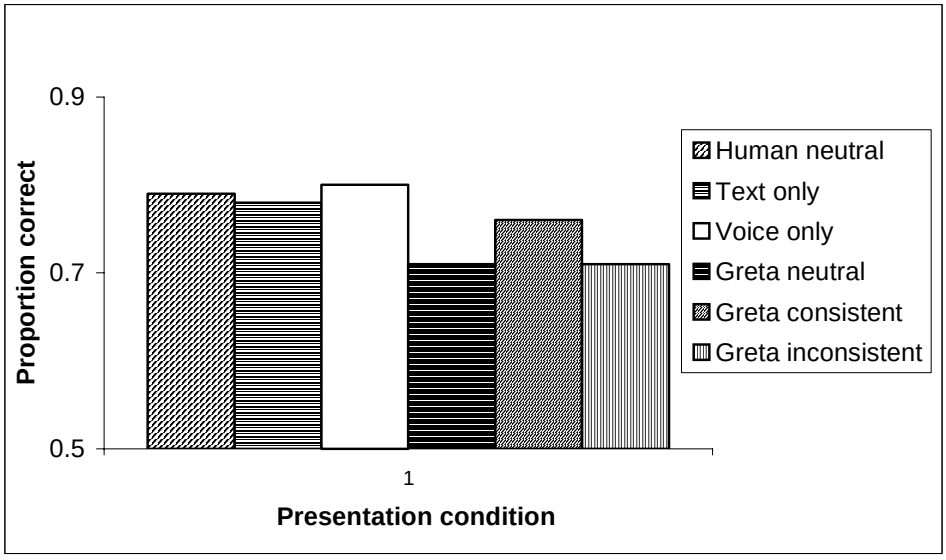
Figure 5: Perceptions of participants towards GRETA, human and voice only conditions

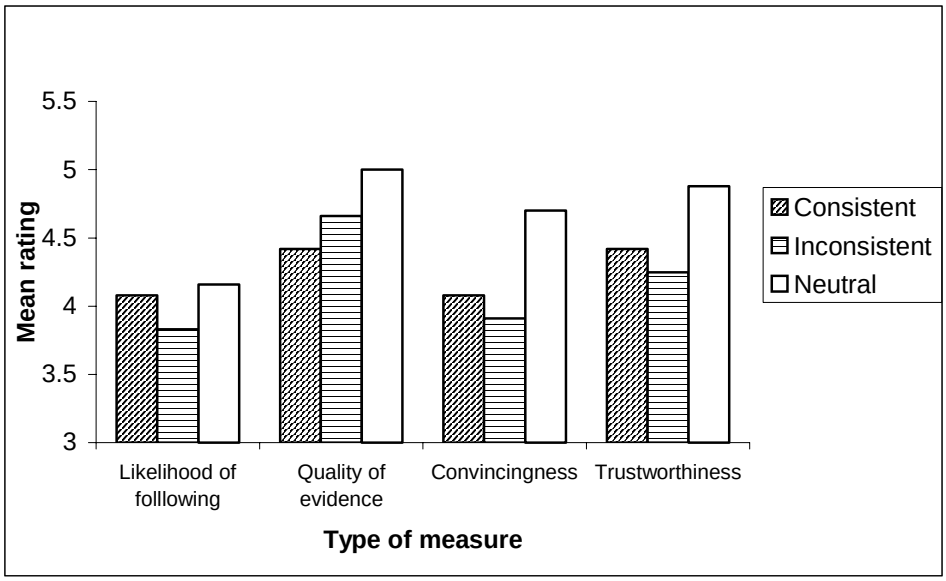
Figure 6: Participants' memory performance for each of the six presentation conditions

Figure 7: Perceptions of participants towards the healthy eating message: comparison of GRETA (neutral), GRETA (consistent), and GRETA (inconsistent) forms of presentation


5.5
5
4.5
4
3.5
3
Mean rating
Likelihood of folllowing
Ease of understanding
Trustworthiness
Type of measure
Greta
Human
Text
Voice


4
3.5
3
2.5
2
Mean rating
Helpful
Likable
Type of measure
Greta
Human
Voice


0.9
0.7
0.5
Proportion correct
1
Presentation condition
Human neutral
Text only
Voice only
Greta neutral
Greta consistent
Greta inconsistent


5.5
5
4.5
4
3.5
3
Mean rating
Likelihood of folllowing
Quality of evidence
Convincingness
Trustworthiness
Type of measure
Consistent
Inconsistent
Neutral

Footnotes

[i] http://www.cstr.ed.ac.uk/projects/festival

[ii] The data for two participants (one each from the voice only and human neutral conditions) was excluded as both were clear outliers (more than 2 standard deviations from the mean).